Summary of Research Activities by Key Approach and Resource

Disease Registries, Databases, and Biomedical Information Systems

Dr. Mark S. Drapkin, Chief, Infectious Disease Service at Newton-Wellesley Hospital (Newton, MA), was treating a 14-year-old girl diagnosed with the fulminant form of meningococcemia. She was delirious and drifting into shock, and death was a real possibility. Specialist colleagues had no suggestions for new treatment approaches. Although the hospital library was closed for the night, Dr. Drapkin had it opened and did a quick Medline search. The search turned up an article in a British journal that suggested plasmapheresis, a procedure designed to remove excess antibodies from the blood by depleting the body of blood plasma without depleting its blood cells. Using the information from the article, Dr. Drapkin and his colleagues successfully treated the patient. "I would never have found these articles in the limited time frame under which we were working without an electronic search of the literature," said Dr. Drapkin.

Introduction

In a world that is increasingly digital, NIH plays a pivotal role in enabling biomedical research, improving health care and public health, and promoting healthy behavior. By connecting and making the results of research—from scientific data to published literature to patient and consumer health information—readily available, NIH magnifies the positive impact of the Nation's investment in the creation of new knowledge in the pursuit of improved health.

Information has become a primary driver of progress in biomedical research and the health care enterprise. For example, genomic data resulting from sequencing the genes of thousands of patients have become primary resources for identifying the genetic basis of diseases. Data that flow from large-scale clinical studies, advanced diagnostic and imaging equipment, and electronic health records are a key enabler of improvements in clinical practice and individual patient care. Up-to-date information from disease registries has become a critical resource for studying disease incidence and treatment patterns, advancing research, and informing public health interventions. The availability of this and other health information available on the Internet offers consumers a more active role in managing their health and further increases demand for reliable and authoritative health information.

The development, deployment, and utilization of disease registries, databases, and other biomedical information systems are essential to managing large amounts of data for research, clinical care, and public health. Such systems permit the efficient collection, organization, storage, sharing, and accessing of biomedical information. Today's biomedical databases house a wide range of clinical, genomic, and other types of scientific data and information resulting from biomedical research and make it accessible for further research or application. Disease registries collect information on cohorts of patients with specific diseases (e.g., cancer, autoimmune disorders, or Parkinson's disease) or who have received specific treatments (e.g., medical devices). They provide a rich source of information that is used by researchers, clinicians, and policymakers.

Increasingly, disease registries and biomedical databases serve not only as repositories of information, but also as research tools in and of themselves, extending and in some cases augmenting the laboratory. Discoveries can be made by examining the information contained in them. For example, scientists can use molecular databases to study the profiles of individual tumors and conceptualize small-molecule anticancer agents to target them. They can analyze large-scale databases linking

genotype and phenotype information from thousands of individuals to identify genes associated with particular observable traits (e.g., obesity) or diseases (e.g., diabetes, cancer). In these ways, biomedical information systems are changing the nature of research itself, and promise to change the nature of clinical care and public health.

The utility of disease registries, databases, and biomedical information systems rests on many factors, including data quality, user accessibility, ease of search capability, availability of useful tools for analysis, and their ability to interoperate with other systems. New data must be added on a regular basis, while existing data are maintained or updated to reflect new findings. Improved search tools are needed to comb through the massive datasets and retrieve relevant results. Standard vocabularies are needed to efficiently organize information, facilitate effective linking and sharing of information, and ensure accurate retrieval, and they, too, must be updated to accommodate new concepts and relationships. New analytical tools are needed to explore increasingly complex questions, such as how the expression patterns of multiple genes are associated with a particular trait or response. Such tools are most effective when information systems are interoperable and can communicate, exchange data, and make use of similar software applications. Given the critical importance of data to biomedical research and health care, policies and procedures are needed to encourage researchers to submit relevant data and to provide other researchers, clinicians, and the general public with suitable access to the data, while simultaneously protecting the confidentiality of personally identifiable information. Preserving, protecting, and ensuring the validity and security of information stored in biomedical databases remains of paramount importance.

Because of the growing importance of information and its management in biomedical science, clinical care, and public health, virtually every NIH IC is engaged in the development, deployment, and use of biomedical information systems that support its mission.

Because of the growing importance of information and its management in biomedical science, clinical care, and public health, virtually every NIH IC is engaged in the development, deployment, and use of biomedical information systems that support its mission. NIH databases and information systems have become indispensable national and international resources for biomedical research and public health. Several trans-NIH activities feature the development of significant biomedical information resources, including the tools, infrastructure, and associated research needed to make databases and registries more valuable. Many of the challenge grants supported with funds from the American Recovery and Reinvestment Act of 2009 (ARRA) relate to data systems and tools, focusing on a wide variety of informatics topics such as new computational and statistical methods for the analysis of large datasets from genome-wide association studies (GWAS) and the use of next-generation sequencing technologies, intelligent search tools for answering clinical questions, and new information technology and resources for disease prevention and personalized medicine.

This section of the Biennial Report describes NIH efforts to develop and deploy disease registries, databases, and biomedical information systems to advance biomedical science, health, and health care. It focuses on:

- Scientific Databases. These databases archive and provide access to authoritative scientific literature, essential research data (including disease-specific data), and clinical research information.
- Genomic Information Systems. Major systems include GenBank for genomic sequence data and dbGaP (database of Genotype and Phenotype) for GWAS data.
- Disease Registries and Surveillance Systems. NIH works with other Federal and private entities to integrate disease registries for national and local use. For example, the Surveillance, Epidemiology, and End Results (SEER) program has been the foundation for innumerable studies, including

recent research into links between hormone therapy and breast cancer.

Emphasis is placed on intramural and extramural activities in which development and maintenance of such information resources is a primary objective, rather than a means of achieving another objective. The described disease registries, databases, and biomedical information systems are intended for widespread use by researchers, clinicians, public health officials, or the general public, and some are associated with policies that require or encourage the submission of particular data or information.

This section of the Biennial Report also describes NIH efforts to make these and other data systems more useful to researchers, clinicians, and the public. Of particular interest are activities related to the following:

- Standardized Vocabularies and Data Protocols. NIH leads the government's efforts to develop standardized vocabularies and terminology to support interoperability among biomedical information systems in research and clinical settings.
- Large-Scale Informatics Infastructure. NIH funds the development of large-scale systems and tools that allow communities of researchers to collect, share, and analyze data needed for research, clinical care (including electronic health records), and public health.
- Biomedical Informatics Research and Training. NIH is the largest Federal funder of biomedical informatics research, which aims to advance the applications of computing to biomedicine for both research and clinical care. Grant programs support research and training in medical informatics and medical librarianship.

Recent developments in policies and procedures to encourage the submission of data to NIH's disease registries, databases, and biomedical information systems also are reviewed.

Catalog of Disease Registry, Database, and Biomedical Information System Activities

In response to the mandate under SEC. 403 (a)(4)(C)(ii) of the Public Health Service Act to provide catalogs of disease registries and other data systems, included here is a live link to an inventory of NIH intramural and extramural activities ongoing in FYs 2008 and 2009 to develop or maintain databases, disease registries, and other information resources for the benefit of the larger research community. Based on a future assessment of the information collected in the inventory, NIH potentially may develop capacity to integrate this catalog as a new category within the NIH RCDC process.

Summary of NIH Activities

NIH Scientific Databases: Enhancing Access to Research Information

Keeping pace with the expanding volume of biomedical knowledge is a continuing challenge for scientists, clinicians, policymakers, and the public; thus, NIH devotes considerable attention and resources to developing, expanding, and maintaining tools and resources for information management. Biomedical databases store and provide access to a wide range of information, from the results of scientific or clinical research studies, to genomic information, to standard reference materials (such as genome sequences or anatomical images), to published journal articles and citations to the medical literature. They are widely used by biomedical researchers, as well as by a growing number of clinicians, public health officials, and consumers. NIH often undertakes special initiatives to make these resources more accessible to a broader, more diverse set of users.

Among the most widely used of NIH's databases are those that collect and provide access to **scientific literature**. These comprehensive resources are extensively used by scientists, health care providers,

and consumers who seek trusted, peer-reviewed information on biomedical and health topics of interest. NIH houses the leading source of authoritative biomedical literature for professional and lay audiences. NIH's exhaustive PubMed/MEDLINE database, for example, indexes citations to articles in more than 5,300 peer-reviewed scientific journals. It contains references to more than 16 million journal articles in the life sciences, and 1.4 million new citations were added to the system during the 2-year period from FY 2008 to FY 2009. During the FY 2008-2009 biennial period, PubMed logged more than 1.5 billion Web-based searches.

The PubMed/MEDLINE database indexes citations to articles in more than 5,300 peer-reviewed scientific journals and contains references to more than 16 million journal articles in the life sciences. Almost 1.4 million new citations were added to the system during the 2-year period from FY 2008 to FY 2009.

In addition, NIH continues to expand PubMed Central (PMC), its digital archive of full-text scientific journal articles. PMC was established to provide online access to a growing number of scientific journal articles deposited by publishers and NIH-funded researchers. Between February 2007 and September 2009, the number of articles available in PMC doubled to 1.9 million and usage rose by more than 60 percent to 360,000 users per day. Some of this increase is attributable to an expanding scope of users—not just biomedical researchers, but also clinicians, other practitioners, and consumers—which highlights the importance of this type of resource.

PMC serves as the repository for manuscripts submitted in accordance with the NIH Public Access Policy, which became mandatory in 2008. The policy ensures that the public and the scientific community have access to the published results of NIH-funded research by requiring NIH-funded scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to PMC. Manuscripts are to be submitted upon acceptance for publication and made accessible to the public no later than 12 months after publication. PMC software also is used by funding agencies in other countries to establish repositories for their funded research. The Wellcome Trust and other major research funders in the United Kingdom established a site that has been operational since 2008, and in 2009, NIH partnered with the Canadian Institutes of Health Research and the National Research Council's Canada Institute for Scientific and Technical Information to establish PMC Canada.

To further facilitate rapid access to emerging scientific findings, NIH announced in August 2009 the introduction of Rapid Research Notes (RRN), a resource to archive research results made available through online venues for rapid scientific communication. The RRN archive allows users to access research that is provided through participating publisher programs for immediate communication. Creation of such an archive had been discussed for many years, but the outbreak of 2009 H1N1 influenza in the spring of 2009 provided increased impetus for the project. The first collection to be archived in RRN will be an open-access, online resource for immediate communication and discussion of new scientific data, analyses, and ideas in the area of influenza. NIH expects the RRN archive to expand over time to include additional collections in other high-interest biomedical fields.

To further facilitate rapid access to emerging scientific findings, NIH announced in August 2009 the introduction of Rapid Research Notes, a resource to archive research results made available through

online venues for rapid scientific communication.

NIH actively endeavors to make its information resources more accessible to varied types of users, as illustrated by its work on MedlinePlus, NIH's comprehensive health information source for consumers and health professionals. Another information source that is directed at a wide variety of users is Genetics Home Reference, NIH's website for consumer-friendly health information on genetic conditions. This information resource bridges consumer health information and scientific bioinformatics data and links to many existing resources at NIH and at other reliable sites.

NIH also puts effort into developing and maintaining information systems that collect data stemming from **biomedical research**, organize it, and make it accessible for subsequent research. NIH's PubChem database, for example, houses data flowing from the high-throughput bioassay centers that were established with NIH funding under the Molecular Libraries Initiative of the NIH Roadmap. It provides information about the biological activity of small molecules, organized as three linked databases along with a chemical structure similarity search tool. The number of unique compounds represented in PubChem more than doubled during FYs 2008-2009 from approximately 10 million to more than 25 million, while the number of bioassays rose from 600 to 1,700. As a result, PubChem provides bioactivity results from more than 50 million tests of small molecules. The number of users per day also increased from approximately 30,000 to 50,000. PubChem is integrated with NIH's Entrez suite of biomedical information resources, enabling users to retrieve related data from multiple databases and navigate among them with relative ease.

NIH is one of three Federal agencies to fund the Protein Data Bank (PDB), an archive of information about experimentally determined structures of proteins, nucleic acids, and complex assemblies. Information on more than 13,000 structures was added to the PDB in FYs 2008-2009, bringing its total content to more than 60,000 molecules. PDB allows users to search for molecules based on their sequence, structure, or function and provides tools to visualize and analyze downloaded structures.

TOXNET is a cluster of 13 large databases covering toxicology, hazardous chemicals, environmental health, and related topics. TOXNET includes literature-based and research databases. It has been used by toxicologists for decades, assisting them in locating toxicology data, literature references, and toxic release information on particular chemicals, as well as in identifying chemicals that cause specific health effects. Peer-reviewed studies from the National Toxicology Program are used by State, local, and Federal health officials to assess the toxicologic potential of environmental compounds to cause adverse health effects such as cancer. To make the Hazardous Substances Data Bank component of TOXNET more useful to first responders at the scene of a disaster, NIH developed WISER, the Wireless Information System for Emergency Responders, which enables wireless access to a selection of the most relevant data for emergency responders. WISER can be installed on personal digital assistants, providing emergency personnel with access to critical information for identifying and safely cleaning up spilled chemicals, understanding their health effects, treating exposed victims, and assessing environmental impact.

To make the Hazardous Substances Data Bank component of TOXNET more useful to first responders at the scene of a disaster, NIH developed WISER, the Wireless Information System for Emergency Responders, which enables wireless access to a selection of the most relevant data for emergency responders.

NIH launched the National Database for Autism Research (NDAR) in FY 2009 as a repository for human subjects data stemming from autism research. NDAR hosts genetic, imaging, and phenotypic research data related to autism and makes it accessible to qualified researchers. The system provides researchers with standards to enable them to analyze and compare data from multiple research sites

and different bioinformatics systems. It also offers bioinformatics tools for depositing, validating, and searching for information. Its collaboration mechanisms allow for sharing quality research data within the autism research community. NIH-funded researchers are strongly encouraged to share their data with NDAR to enable secondary use and analysis.

Another group of NIH-supported databases organize and provide access to **clinical research information**. NIH's ClinicalTrials.gov database was significantly enhanced during FYs 2008 and 2009 to respond to the Food and Drug Administration Amendments Act of 2007 (Pub. L. No. 110-85), which expands the types of clinical trials that must be registered in ClinicalTrials.gov, increases the amount of information that must be submitted for each trial, and requires the submission of summary results data, including adverse events. During FYs 2008 and 2009, more than 34,000 trials were registered with ClinicalTrials.gov, raising the total number of registered trials in the system to 80,000. During that same time period, summary results of more than 830 trials were submitted to the system and made available to the research community and the general public.

In addition, the NIH Biomedical Translational Research Information System (BTRIS), which was initiated in 2008, was made available to the intramural NIH community in 2009. BTRIS is a powerful new tool for NIH investigators to access clinical research data, develop streamlined mechanisms for protocol reporting and data analysis, and reuse data for hypothesis generation and collaboration. New functionality will continue to be added to the system.

ProtoType is an assisted protocol authoring tool that provides a systematic framework where research protocols can be developed and maintained throughout their life cycle. ProtoType includes fully customized documents tailored toward individual Institutional Review Boards, allowing investigators to focus on the substance of their protocols, rather than the formatting.

Genomic Information Systems: Understanding the Genetic Basis of Disease

NIH also has made great strides in developing information resources to support genetics research. NIH has long supported genetics research through widely used resources such as GenBank, the NIH genetic sequence database. In FY 2009, NIH launched the Sequence Read Archive (SRA) to accommodate the massive quantities of data coming from sequencing projects that are using new high-throughput technologies. SRA is proving to be one of the fastest growing biological databases in history, with more than 10 terabytes of sequence data under management at the end of FY 2009 and a growth rate of about 1 terabyte per month. NIH's Influenza Virus Resource database, comprising information obtained from the NIH Influenza Genome Sequencing Project and GenBank, contains more than 90,000 influenza virus sequences, including the sequences of more than 2,000 whole influenza genomes. In spring 2009, with the rapid emergence of the 2009 H1N1 pandemic, the database received more than 2,200 influenza sequences from the Centers for Disease Control and Prevention and laboratories from 35 countries. This resource enables scientists to compare influenza virus strains so that emergent variants can be identified more rapidly and vaccines developed accordingly. As the library of viral sequences grows, it will be an increasingly important reference to help further understand how avian viruses spread to humans, and how influenza activity spreads throughout the world.

Considerable effort has been aimed at supporting the analysis of data from GWAS, which explore the connection between specific genes (genotype information) and observable diseases or conditions (phenotype information, such as diabetes, high blood pressure, or obesity). NIH's dbGaP (database of Genotype and Phenotype) houses data from a number of GWAS, including those funded by NIH. By the end of 2009, dbGaP included results from more than 40 GWAS, including genetic analyses related to such diseases as Parkinson's disease, amyotrophic lateral sclerosis, diabetes, alcoholism, lung cancer, and Alzheimer's disease. NIH's GWAS policy, which went into effect in January 2008, encourages NIH grantees to submit their GWAS data to dbGaP and establishes procedures for making it available to other researchers to speed up disease gene discovery while at the same time protecting

the privacy of research subjects in genomics studies.

By the end of 2009, dbGaP (database of Genotype and Phenotype) included results from more than 40 genome-wide association studies, including genetic analyses related to such diseases as Parkinson's disease, amyotrophic lateral sclerosis, diabetes, alcoholism, lung cancer, and Alzheimer's disease.

In addition, several NIH ICs have established genetics repositories to accelerate research and multidisciplinary collaborations in specific disease areas. Programs such as the NEI eyeGENE, NIMH Genetics Repository, the NINDS Human Genetics Repository, the NIEHS Chemical Effects in Biological Systems (CEBS) Knowledge Base, and the NIA Genetics of Alzheimer's Disease Data Storage Site give researchers access to vast storehouses of genetic and genomic data, DNA samples, and clinical data, along with informatics tools designed to facilitate their analyses. The wide availability of information linking genotype to phenotype should help researchers better understand gene-based diseases and speed development of effective therapies.

Other NIH-supported genetic databases contain information on model organisms, which are widely used by researchers to understand disease processes and develop new therapeutic strategies and tools that can be transferred to humans. The NIH-funded Rat Genome Database, for example, combines information on the genome, genes, and disease traits of different strains of rats with related information on the mouse and human genomes. WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics, and biology of *C. elegans* and related organisms. The Universal Protein Resource (UniProt) Knowledgebase offers the scientific community free access to a comprehensive source of information on protein sequences and related functional information.

Disease Registries and Surveillance Systems: Tracking and Monitoring Disease

Disease registries collect information about the occurrence of specific diseases, such as cancer and Parkinson's disease, the kinds of treatment that patients receive, and other information that might be relevant to researchers or public health officials. Increasingly, disease registries also include genomic data from registered patients. Registry information can therefore help identify causal factors of disease, assess the effectiveness of various interventions, and identify questions of concern to researchers, clinical professionals, and policymakers.

NIH-supported disease registries have paid many dividends over the years. Recently, for example, with the participation of patients from the Alopecia Areata Registry, NIH-supported scientists discovered four chromosomal locations that appear to be associated with susceptibility to this common autoimmune disease, which is characterized by patchy hair loss. Understanding the mechanisms of the genes found at these locations could lead to the development of an effective treatment for the disease, which is presently untreatable.

Disease registries have been employed for research on other autoimmune disorders, including Sjogren's Syndrome, one of the most prevalent. A significant roadblock for moving discoveries ahead in the field of Sjogren's Syndrome is a lack of data and biospecimens available for research. Recognizing the problem, NIH spearheaded an effort to establish patient registries at two extramural institutions, as well as through its own intramural program. These groups work together to generate and share genome-wide genotyping data and clinical information from the cohorts enrolled through these efforts with the general research community. Similarly, the International Epidemiologic Databases to Evaluate AIDS (IeDEA) aims to establish centers in multiple regions of the world for the collection and harmonization of data that can be used by an international research consortium to address unique and evolving research questions in HIV/AIDS that are currently unanswerable by single cohorts. High-

quality data are being collected by researchers throughout the world. This initiative provides a means to establish and implement methodology to effectively pool the collected data—thus providing a cost-effective means of generating large datasets to address high-priority research questions.

The inclusion of genomic information in disease registries makes them valuable resources for investigating the contribution of genes and genetic variation to diseases of interest. To spur such research, NIH collaborated with the University of North Carolina's General Clinical Research Center to launch a large volunteer DNA banking project named the Environmental Polymorphisms Registry (EPR), which will collect DNA samples from up to 20,000 individuals in the greater North Carolina Triangle Region. These samples will be available to scientists to look for genes that may be linked to common diseases such as diabetes, heart disease, cancer, asthma, and many others. In addition, NIH supports the National Registry of Genetically Triggered Thoracic Aortic Aneurysms and Cardiovascular Conditions (GenTAC). The goal of GenTAC is to establish a registry of patients with genetic conditions that may be related to thoracic aortic aneurysms—a disorder that weakens the main artery from the heart—and to collect medical data and biologic samples. The samples and data are made available to qualified investigators to enable research on effective medical practices and to advance the clinical management of genetic thoracic aortic aneurysms, and other cardiovascular complications. NIH supports several other registries associated with specific diseases, including lupus, muscular dystrophy, and rheumatoid arthritis.

The International Epidemiologic Databases to Evaluate AIDS aims to establish centers in multiple regions of the world for the collection and harmonization of data that can be used by an international research consortium to address unique and evolving research questions in HIV/AIDS that are currently unanswerable by single cohorts.

Registries also serve as an effective mechanism to gather data on the incidence, prevalence, and natural history of diseases. The NIH-supported California Parkinson's Disease Registry, for example, enables researchers to identify the possible environmental and genetic origins of this progressive neurological disorder suffered by an estimated 1.5 million Americans. Data in the registry can help to determine whether race, ethnicity, gender, age, environmental factors, or place of residence influence the likelihood of getting the disease, and can help track incidence and demographic trends.

Registries also provide a valuable source of information for tracking the effectiveness of particular treatments or interventions. The Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS), for example, is a national registry for patients who are receiving mechanical circulatory support device therapy to treat advanced heart failure. The registry is supported jointly by NIH, the Food and Drug Administration, and the Centers for Medicare and Medicaid Services. Use of standardized terminologies helps ensure that the data collected will facilitate improved patient evaluation and management while aiding in better device development. INTERMACS also is expected to facilitate appropriate regulation and reimbursement of the implantation of mechanical circulatory support devices.

Registries also are integral elements of more comprehensive NIH programs designed to monitor and analyze disease trends in the United States. For example, the SEER program has a 35-year track record of identifying emerging trends, geographic variation, ethnic disparities, and other patterns that have provided new directions for epidemiologic research into the cause, progression, and control of cancer. SEER collects and publishes cancer incidence and survival data from cancer registries covering approximately 26 percent of the American population. De-identified data is made available for research, and an interactive query system is available on its website. SEER data provided critical insight into the relationship between hormone therapy and breast cancer incident rates. SEER data recently have been enhanced by linking persons in SEER to Medicare enrollment and utilization data. The SEER-Medicare data are longitudinal and can be used to assess health care received prior to a

cancer diagnosis, at the time of diagnosis, and after initial treatment until death. There have been more than 400 peer-reviewed publications resulting from SEER-Medicare data, adding to the thousands of publications based on SEER.

Surveillance and monitoring programs also are crucial sources of information and analysis for policymakers, legislators, public health officials, clinicians, and the public. SEER participates in Cancer Control P.L.A.N.E.T., a Web portal that provides links to comprehensive cancer control resources and data for public health professionals. NIH supports several epidemiologic programs designed to gather ongoing data and monitor emerging drug abuse trends in adolescents and other populations, helping to guide national and global prevention efforts, drug control, and public health policy. Among the projects are the Monitoring the Future (MTF) Survey, which has been tracking trends in substance use, attitudes, and beliefs among adolescents and young adults in the United States since 1975, and the Community Epidemiology Work Group (CEWG), which provides ongoing community-level surveillance of drug abuse through analysis of quantitative and qualitative research data. CEWG findings reported in 2008 and 2009 show decreases in methamphetamine indicators (e.g., treatment admissions), suggesting that the problems that had escalated in the first half of the decade may have stabilized or declined.

NIH supports several epidemiologic programs designed to gather ongoing data and monitor emerging drug abuse trends in adolescents and other populations, helping to guide national and global prevention efforts, drug control, and public health policy.

NIH also supports the Alcohol Policy Information System (APIS), an online database that provides detailed information on a wide variety of alcohol-related policies in the United States at both State and Federal levels. Designed primarily as a tool to encourage and facilitate research on the effects and effectiveness of alcohol-related public policies in the United States, APIS simplifies the process of ascertaining the state of the law for studies on the effects and effectiveness of alcohol-related policies.

Standardized Vocabularies, Data Protocols, and Tools

NIH continues to invest in tools that can increase the utility of its scientific databases and medical information sources. A key component of such efforts relates to the development and maintenance of standards and vocabularies for use in information systems used for research and clinical care, including electronic health records. Medical terminology can be difficult to remember and can vary from one laboratory or clinical facility to another. Often there are many names for a single concept (e.g., cancer of the colon, colonic neoplasm, colon cancer). Standard vocabularies and ontologies (models of the relationships between concepts) improve information search, retrieval, and exchange by endowing systems with the ability to automatically perceive and retrieve information about related terms. As expansion of scientific frontiers produces new concepts, terms, and relationships, standard vocabularies must be regularly revised so that articles and other data can be properly indexed and search engines can find relevant and related terms.

NIH continues to update the Unified Medical Language System (UMLS), which is used heavily in advanced biomedical research and data mining worldwide. The UMLS Metathesaurus, with more than 7.7 million concept names from more than 100 vocabularies, is a distribution mechanism for standard code sets and vocabularies used in health data systems. Many institutions apply UMLS resources in a wide variety of applications including information retrieval, natural language processing, creation of patient and research data, and the development of enterprise-wide vocabulary services for electronic health records.

NIH is pursuing research and development on robust and scalable approaches to synthesizing,

representing, updating, and deploying electronic knowledge and decision algorithms for use in conjunction with electronic health records.

The broad deployment and use of advanced electronic health records will provide expanded opportunities for access to biomedical knowledge and advanced decision support for the public, their health care providers, and the public health workforce. To turn this potential into effective reality, NIH is pursuing research and development on robust and scalable approaches to synthesizing, representing, updating, and deploying electronic knowledge and decision algorithms for use in conjunction with electronic health records. NLM serves as the central coordinating body for clinical terminology standards across the HHS and works closely with the Office of the National Coordinator for Health Information Technology (ONC) to support nationwide implementation of an interoperable health information technology infrastructure. NIH develops and licenses key clinical terminologies that are designated as standards for health information exchange in the United States. It produces RxNorm, a standard clinical drug vocabulary, supports the Logical Observation Identifiers Names and Codes (LOINC) nomenclature for laboratory tests and patient observations, and collaborates with the International Health Terminology Standards Development Organisation to promote international adoption of the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). In FY 2009, NIH released the first version of the CORE Problem List Subset of SNOMED CT, designed to facilitate coding of problem list data in electronic health records by mapping frequently used terms from seven large-scale health care institutions to corresponding SNOMED CT concepts. (The problem list is often the first part of the clinical narrative in an electronic health record that is codified with some controlled vocabulary.) The Newborn Screening Codes and Terminology Guide, a Web portal to support more effective use of newborn screening laboratory test information, was created in FY 2009 in collaboration with the ONC, the Health Resources and Services Administration, and newborn screening organizations.

Common terminologies are a key enabler of related research and development to exploit the inherent relationships among information in disparate databases and support the interlinking of data systems. PubChem's chemical structure and bioassay records, for example, are interlinked with the biomedical literature in PubMed and with three-dimensional protein structure records. This integration provides many routes by which biomedical researchers may discover the candidate probes developed by the Molecular Libraries Initiative. A researcher examining a protein sequence record, for example, may see that a particular protein has been screened, view the active compounds, and examine structure-activity relationships using PubChem analysis tools. Another NIH resource, the Daily Med, is an official distribution mechanism for FDA-approved packaging information (drug label inserts) that links to other sources of drug information, including MedlinePlus, ClinicalTrials.gov, and PubMed. More than 60,000 people subscribe to its RSS data feeds.

NIH's Discovery Initiative, launched in FY 2006-2007 and continuing into FY 2008-2009 aims to take database linking to the next level. The Discovery Initiative will improve the presentation of results from search queries conducted across a range of NIH databases so that users, who often do not go beyond retrieving the basic results of a search query, are more likely to be drawn to related information that could lead to serendipitous discoveries, even if that information resides in another NIH database. NIH's Collective Intelligence Initiative aims to facilitate data re-use and knowledge discovery by using controlled vocabularies and ontologies to pull together and analyze related information from databases across the ICs.

Large-Scale Informatics Infrastructure

NIH also has embarked on a number of large-scale initiatives to develop and deploy infrastructure and tools for storing, sharing, integrating, and analyzing the large volumes of data routinely generated in research laboratories and in clinical settings. These initiatives tend to produce not only storehouses for data generated by research, but also larger scale networks for sharing data, linking researchers, and

conducting further research. NIH supports a number of clinical research networks, for example, infrastructure that allows standardized data reporting and sharing of information across clinical studies. (Also see the section on *Clinical and Translational Research* in Chapter 3.)

In the area of cancer research, NIH has established the Cancer Biomedical Informatics Grid® (caBIG®), a collaborative information network for all of NCI's advanced technology and program initiatives that aims to enable collaborative research and personalized, evidence-based care. The network connects scientists, practitioners, and patients, enabling the collection, analysis, and sharing of data and knowledge along the entire research pathway from bench to bedside. Specific biomedical research tools under development by caBIG® include clinical trial management systems, tissue repositories and pathology tools, imaging tools, and a rich collection of integrative cancer research applications. Ongoing collaborations with research and bioinformation organizations in China, India, and the United Kingdom are driving international adoption of caBIG® resources. The caBIG® infrastructure also supports a new health care ecosystem, BIG Health, launched in 2008 in collaboration with various stakeholders in biomedicine (e.g., government, academia, industry, nonprofits, and consumers) in a novel organizational framework to demonstrate the feasibility and benefits of personalized medicine. BIG Health will provide the foundation for a new approach in which clinical care, clinical research, and scientific discovery are linked.

Other efforts aim to provide the informatics infrastructure to advance basic research and clinical studies across the spectrum of biomedical sciences. NIH's Biomedical Informatics Research Network (BIRN) is a virtual community of shared informatics resources. BIRN's grid computing technology makes digital research data freely available for sharing and exchange among communities of researchers; its data integration tools allow searching across distributed databases; and it provides tools for data analysis, management, and collaborative research. (Also see the section on *Technology Development* in Chapter 3.) The CardioVascular Research Grid (CVRG) provides infrastructure for sharing cardiovascular data and data analysis tools. The CVRG builds on and extends tools developed in the caBIG® and BIRN projects to support national and international collaborations in cardiovascular science.

A Disaster Information Management Research Center was established in FY 2008 to facilitate access to disaster information, promote more effective use of libraries and disaster information specialists in disaster management efforts, and ensure uninterrupted access to critical health information resources when disasters occur.

The Neuroscience Information Framework (NIF), part of NIH's Neuroscience Blueprint, aims to advance neuroscience research by enabling discovery and access to research data and tools worldwide through an open source, networked environment. By the end of FY 2009, more than 2,300 Web-accessible information resources were listed in the NIF registry. NIF also supports the NeuroLex project, which aims to define neurological terms and their relationships to simplify information retrieval and sharing. NeuroLex consisted of approximately 17,000 neuroscience concepts at the end of FY 2009. A related effort, the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC), facilitates finding and comparing structural and functional neuroimaging tools and resources. Collecting and pointing to standardized information about tools, this site helps researchers find the right structural or functional neuroimaging tool or resource and determine whether it can contribute to a given research endeavor. More than half the tools available through the Clearinghouse previously were unavailable for sharing. Since its release at the beginning of FY 2008, more than 50,000 software files have been downloaded from its award-winning website.

The Bioinformatics and Computational Biology initiatives of the NIH Roadmap continue to make progress toward creation of a national biomedical data and information management system. Through the system, biologists, chemists, physicists, computer scientists, and physicians anywhere in the country will be able to use a common set of software tools to analyze, integrate, model, simulate, and

share data. The National Centers for Biomedical Computing are a central focus of this effort, providing funding for seven centers that cover systems biology, image processing, biophysical modeling, biomedical ontologies, information integration, and tools for gene-phenotype and disease analysis. The Centers collaborate with other NIH-funded institutions on topics ranging from biomechanics to standards development for data mining, and cross-Center working groups pursue activities of common interest, such as Biositemaps, which assist users in locating, querying, composing or combining, and mining biomedical information from databases that are distributed across the Centers.

NIH also develops advanced information infrastructure to assist emergency responders when disaster strikes. A Disaster Information Management Research Center was established in FY 2008 with the aim to facilitate access to disaster information, promote more effective use of libraries and disaster information specialists in disaster management efforts, and ensure uninterrupted access to critical health information resources when disasters occur. A disaster information website provides access to a broad range of emergency preparedness and response information. The Center also collaborates with the Navy National Medical Center, Suburban Hospital, Johns Hopkins Medicine, and the NIH Clinical Center through the Bethesda Hospital Emergency Preparedness Partnership. The Partnership provides backup communication systems and develops tools for patient tracking, information sharing and access, and responder training, serving as a model for hospitals across the Nation.

Biomedical Informatics Research and Training

Ensuring continued advances in biomedical informatics resources requires active support of fundamental research that seeds the further development of new tools, resources, and approaches. It also is critical to generate a continuous supply of skilled biomedical informatics researchers, information specialists (such as medical librarians), and life sciences researchers trained in bioinformatics. NIH continues to expand its efforts in bioinformatics research and training in response to the growing importance of informatics in the biomedical and life sciences.

NIH supports research in new technologies to address issues such as: interoperability of data systems, compatibility of computer software across medical institutions, security of data during transmission, compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA), availability of affordable data systems for patient care providers, and integration of medical decision support information in medical data systems. Several ICs fund informatics research projects within their areas of specialization. NLM remains the primary Federal sponsor of biomedical informatics research, and its extramural grants program supports research on the characterization, management, and efficient use of data, information, and knowledge in health care and basic biomedical sciences. Grants funded in FY 2008-2009 explored informatics challenges related to clinical care, biomedical research, genomics, and public health. NLM's most recent long-range plan, *Charting a Course for the 21st Century,* identifies a number of emerging informatics challenges that will demand continued research and development.

Funds from the American Reinvestment and Recovery Act will allow NIH to support an additional 56 2-year slots at 10 of its training programs for 2009 and 2010.

NIH also is the principal source of support for research training in biomedical informatics, providing research training grants to 18 institutions that enrolled 270 trainees in FY 2009. ARRA funds will allow NIH to support an additional 56 2-year slots at 10 of its training programs for 2009 and 2010. NIH also implemented a Diversity Short-Term Trainee Program in FY 2008 that supported 18 trainees in 7 training programs and sponsors an Informatics Training for Global Health Program, which supports informatics research training in low- and middle-income country institutions in partnership with U.S. institutions and investigators. Training is integrated with ongoing research at the foreign institutions to develop informatics capacity and support research. Training must address the health and informatics needs of the collaborating countries. (Also see the section on *Research Training and Career*

Development in Chapter 3.)

Conclusion

The results of NIH's commitment to disease registries, databases, and biomedical information systems are apparent in the following highlights describing some of the important accomplishments and ongoing initiatives.

Notable Examples of NIH Activity

Key
E = Supported through **E**xtramural research I = Supported through **I**ntramural research O = **O**ther (e.g., policy, planning, or communication) COE = Supported via congressionally mandated **C**enter **o**f **E**xcellence program GPRA Goal = **G**overnment **P**erformance and **R**esults **A**ct ARRA = **A**merican **R**ecovery and **R**einvestment **A**ct IC acronyms in **bold** face indicate lead IC(s).

NIH Scientific Databases: Enhancing Access to Research Information

PubMed®/MEDLINE®: NIH continued to expand PubMed/MEDLINE as a tool for biomedical research, clinical medicine, and consumer health. Nearly 1.4 million articles from the biomedical journal literature were added to PubMed/MEDLINE in FYs 2008-2009. The Indexing 2015 initiative continues to pursue increases in the speed and efficiency of indexing through natural language processing and other automated techniques, and in FYs 2008-2009, the GPRA goal of reducing the time to catalog new journals added to NLM's collection was achieved. Drawing on results of the NCBI Discovery Initiative, enhancements were made to PubMed search capabilities to expand the number and highlight the visibility of links to related information across multiple databases.

- For more information, see http://pubmed.gov
- (I) (**NLM**) (GPRA)

PubMed Central (PMC): NIH made significant enhancements to PMC, an online repository of full-text biomedical journal articles. Since February 2007, PMC has doubled the number of articles to 1.9 million in September 2009. Usage also has risen by 60 percent to more than 360,000 users per day. A part of the growth has stemmed from the NIH Public Access policy changing from a voluntary to a mandatory program in April 2008. Under the Public Access policy, all NIH-funded research articles must be deposited in PMC. The NIH Manuscript Submission system streamlines the process for NIH-funded authors submitting their manuscripts to PMC, and more than 5,000 manuscripts a month are received, compared to less than 1,000 under the previous voluntary Public Access policy in 2007. Through PMC agreements with publishers, a growing number of journals, now more than 700, offer full text of their contents to PMC either immediately upon publication or within 12 months. To foster international cooperation on preservation and access to biomedical literature, NIH made PMC software available to archiving organizations outside the United States and worked with the Wellcome Trust and other major United Kingdom (UK) research funders to establish a collaborating PMC site in the UK, which has been operational since 2008. In 2009, NIH partnered with the Canadian Institutes of Health Research and the

National Research Council's Canada Institute for Scientific and Technical Information to establish PMC Canada.

- For more information, see http://www.pubmedcentral.nih.gov
- For more information, see http://ukpmc.ac.uk
- (I) (**NLM**)

PubChem: PubChem is an open repository for data on the properties of small molecules, including bioactivity test results. It began in 2004 as part of NIH's Molecular Libraries Program, which aims to discover new chemical probes through high-throughput biological screening. As of FY 2009, there were more than 25 million unique structures and 1,700 bioassays. These assays contain information on the biological activities of 700,000 compounds, yielding more than 50 million bioactivity results, and have been contributed by 34 academic, government, and commercial organizations. Through the PubChem website, more than 50,000 scientists a day rapidly search chemical structures, retrieve and compare screening results, explore structure-activity relationships, and identify potential molecular targets.

- Wang Y, et al. *Nucleic Acids Res* 2009;37(Web Server issue):W623-33. PMID: 19498078. PMCID: PMC2703903.
- For more information, see http://pubchem.ncbi.nlm.nih.gov/
- (I) (**NLM**)

TOXicology Data NETwork (TOXNET): TOXNET is a cluster of 13 databases covering toxicology, hazardous chemicals, environmental health, and related topics. It is a primary reference for toxicologists, poison control centers, public health administrators, physicians, and other environmental health professionals, and includes databases such as Hazardous Substances Data Bank, TOXLINE, GENE-TOX, and the Toxic Release Inventory. In FY 2008, the Carcinogenic Potency Database at the University of California, Berkeley, which reports analyses of animal cancer tests and is in support of cancer risk assessments, was added to the databases searchable through the TOXNET search engine. TOXNET is highly used, with nearly 600,000 users in FYs 2008 and 2009. Enhancements based on user feedback were made in FY 2008.

- (I) (**NLM**)

National Database for Autism Research: The National Database for Autism Research (NDAR) is a collaborative biomedical informatics system created by NIH to provide a national resource to support and accelerate research in autism spectrum disorder (ASD). NDAR hosts human genetic, imaging, and phenotypic research data relevant to ASD, making these data available to qualified researchers. NDAR also has the capability to allow investigators to use NDAR for data sharing among select collaborators in ongoing studies. Through its Data Dictionary, NDAR will foster the development of a shared, common understanding of the complex data landscape that characterizes ASD research. Finally, its architecture facilitates linkage of NDAR with other significant data resources, regardless of their location or ownership and in ways that respect the policies and implementations of those other data resources.

- For more information, see http://ndar.nih.gov/
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System* and Chapter 2: *Life Stages, Human Development, and Rehabilitation*
- (E/I) (**NIMH**, CIT, NICHD, NIDCD, NIEHS, NINDS)

National NeuroAIDS Tissue Consortium: The National NeuroAIDS Tissue Consortium (NNTC) is a repository of brain tissue and fluids from highly characterized HIV-positive individuals. Established as a resource for the research community, the NNTC includes information from more than 2,280 participants in its clinical evaluation/tissue donation program, including nearly 750 brains, thousands of plasma and cerebrospinal fluid samples, and additional organs and nerves of interest.

- For more information, see http://www.hivbrainbanks.org/
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System*
- (E/I) (**NIMH**, NINDS)

ClinicalTrials.gov: ClinicalTrials.gov was significantly modified during FY 2008-2009 to respond to new clinical trial registration and results reporting requirements established by the FDA Amendments Act of 2007 (PL 110-85). The existing registry was expanded to accommodate the submission of more information about a larger number of trials, including those trials of FDA-regulated drugs, biological products and devices that now are required to register. In addition, NIH developed and implemented results modules to accept and display to the public summary results information, including adverse event information from registered trials. Mandatory reporting of results began in September 2008, with mandatory submission of adverse event information following in September 2009. During FYs 2008-2009, more than 34,000 trials were newly registered with ClinicalTrials.gov, raising the total number of registered trials to 80,000. In addition, summary results of more than 830 clinical trials were submitted and made available at ClinicalTrials.gov, with the rate of results submission approaching 200 trials per month by the end of FY 2009. To solicit input on issues to be considered in rulemaking for further expansion of ClinicalTrials.gov, a public meeting was held in April 2009; more than 200 participants attended the meeting, and more than 70 written comments were submitted to a public docket.

- This example also appears in Chapter 3: *Clinical and Translational Research*
- (I) (**NLM**)

The Biomedical Translational Research Information System: The NIH intramural program uses a wide variety of clinical and research data management systems to gather clinical research data. The single largest system is the Clinical Research Information System (CRIS) at the NIH CC; however, many of the other 26 ICs at NIH have their own systems, as do many laboratories at the ICs and even individual researchers within the laboratories. Thus, research data for individual clinical trials and on individual subjects are scattered across multiple diverse systems. The Biomedical Translational Research Information System (BTRIS) project, initiated in 2008 and made available in 2009, includes a sophisticated data warehouse that currently contains the data on more than 447,000 subjects from 8,800 protocols, gathered from the CRIS system (2004 to present), archived data from CRIS' predecessor system (1976 to 2004), and data from systems at NIAID and NIAAA. BTRIS provides NIH researchers with a user-friendly reporting application to obtain data on subjects in their own protocols from across all these sources. They also are able to perform queries against all data on all research subjects from all clinical trials (in de-identified form), to allow them to ask new questions of the data, look for previously unrecognized correlations, and gain new insights through the reuse of data that NIH has been collecting for the past three decades.

- For more information, see http://btris.nih.gov
- (I) (**CC**)

ProtoType: ProtoType is an assisted protocol authoring tool that provides a systematic framework where protocols can be developed and maintained throughout their life cycle. ProtoType includes fully customized documents tailored toward individual Institutional Review Boards (IRBs), taking all the guesswork out of creating a protocol and allowing the investigator to focus on authoring. By capturing the entire authoring process electronically, the protocol can be moved easily between the IC IRB, NIH CC, and other Institutes and investigators while tracking the state of the protocol. Ultimately, ProtoType will be an online archive of all protocols submitted by each principal investigator and will maintain the protocols' histories. Prototype currently has incorporated templates from 4 of the 12 NIH IRBs, with more than 200 distinct investigators using the system.

- For more information, see https://prototype.cc.nih.gov
- (I) (**CC**)

Genomic Information Systems: Understanding the Genetic Basis of Disease

Influenza Virus Resources: NIH maintains the Influenza Virus Resource, a database of influenza virus sequences that enables researchers around the world to compare different virus strains, identify genetic factors that determine the virulence of virus strains, and look for new therapeutic, diagnostic, and vaccine targets. The resource was developed using publicly accessible data from laboratories worldwide in addition to targeted sequencing programs such as NIH's Influenza Genome Sequencing Project. Updated daily, this comprehensive sequence resource includes more than 90,000 influenza sequences and more than 2,000 complete genomes. In the spring of 2009, with the rapid emergence of the 2009 H1N1 pandemic, the database received more than 2,200 influenza sequences from publicly accessible databases and included sequences from CDC and labs from 35 countries. By the end of 2009, nearly 10,000 H1N1 sequences were in the database. The combination of extensive sequence data and advanced analytic tools provided researchers worldwide immediate access for investigating the rapid spread of this flu and developing vaccines for combating it. Other influenza virus information resources also were developed in response to 2009 H1N1. To facilitate access to the scientific literature, a pre-formulated search for 2009 H1N1 papers was added to PubMed. A 2009 H1N1 Flu page with comprehensive information on Federal response, international resources, transmission, prevention, treatment, genetic makeup, and veterinary resources was added to Enviro-Health Links, which provides links to toxicology and environmental health topics of recent special interest, including information in Spanish. For the general public, patients, family members, and caregivers, a health topic on 2009 H1N1 flu, in Spanish and English, was added to the MedlinePlus consumer health resource.

- Bao Y, et al. *J Virol* 2008;82(2):596-601. PMID: 17942553. PMCID: PMC2224563.
- For more information, see http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
- For more information, see http://www.pubmed.gov
- For more information, see http://sis.nlm.nih.gov/enviro/swineflu.html
- For more information, see http://www.nlm.nih.gov/medlineplus/h1n1fluswineflu.html
- For more information, see http://www.nlm.nih.gov/medlineplus/spanish/h1n1fluswineflu.html
- This example also appears in Chapter 2: *Infectious Diseases and Biodefense* and Chapter 3: *Molecular Biology and Basic Research*
- (I) (**NLM**)

Genome-Wide Association Studies: With unprecedented speed, researchers have used an approach called genome-wide association studies (GWAS) to explore genetic variants and their complex relationships to human health and disease. GWAS research has linked a stunning number of

genetic variants to common conditions—more than 130 in 2008 alone. For example, the obesity epidemic and its related health conditions pose a great challenge for the Nation. In 2008, the Genetic Investigation of Anthropometric Traits consortium identified six genes associated with body mass index, a key indicator for obesity. Also in 2008, three GWAS of lung cancer implicated several genes already known to be linked to nicotine addiction. In a feat that would not have been possible without the power of whole genome analysis, the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium in 2009 gathered data from participants in long-running studies to reveal genetic variants associated with an increased risk of stroke. Identification of genetic variants associated with common diseases opens new windows into the biology of health and disease. This work also raises the possibility of someday using genetic testing, in combination with family history, to identify at-risk, pre-symptomatic individuals who might benefit from personalized screening and preventive therapies.

- For more information, see http://www.genome.gov/27528559
- For more information, see http://www.genome.gov/27529231
- For more information, see http://www.genome.gov/27531390
- This example also appears in Chapter 2: *Cancer*, Chapter 2: *Chronic Diseases and Organ Systems* and Chapter 3: *Genomics*
- (E, I) (**NHGRI, NIDDK, NCI, NIA, NHLBI, NIMH**, NINDS)

Database of Genotype and Phenotype (dbGaP): Research on the connection between genetics and human health and disease has grown exponentially since completion of the Human Genome Project in 2003, generating high volumes of data. Building on its established research resources in genetics, genomics, and other scientific data, NIH established dbGAP to house the results of genome-wide association studies (GWAS), which examine genetic data of de-identified subjects with and without a disease or specific trait to identify potentially causative genes. By the end of 2009, dbGaP included results from more than 40 GWAS, including genetic analyses related to such diseases as Parkinson's disease, ALS, diabetes, alcoholism, lung cancer, and Alzheimer's disease. dbGaP is the central repository for many NIH-funded GWAS to provide for rapid and widespread distribution of such data to researchers and accelerate the understanding of how genes affect the susceptibility to and severity of disease.

- For more information, see http://view.ncbi.nlm.nih.gov/dbgap
- This example also appears in Chapter 3: *Epidemiological and Longitudinal Studies* and Chapter 3: *Genomics*
- (I) (**NLM**)

Genome-Wide Association Studies of Autoimmune Disease Risk: In recent years, genome-wide association studies (GWAS) have transformed the identification of gene regions related to disease risk, through an unbiased analysis of patients with a disease, in comparison with people who don't have it. These GWAS require large numbers of patients and individuals without the disease to obtain statistically significant results. Long-term NIH support of disease registries and repositories of biological samples have been essential to successful projects, in addition to productive, multisite collaborations across the United States, including international researchers and contributions from the NIH Intramural Research Program. GWAS have yielded important information about disease risk, as well as understanding of disease pathways and potential therapeutic targets, in several autoimmune diseases in the past 2 years. Diseases studied include psoriasis, rheumatoid arthritis, systemic lupus erythematosus (or lupus), ankylosing spondylitis, and type 1 diabetes. Initial results from GWAS require confirmation by replication in additional groups of patients. More detailed localization of disease risk genes can be achieved through comprehensive DNA sequencing of candidate gene regions. New NIH

initiatives are supporting these follow-up studies, which are critical to validating GWAS findings.

- Plenge RM, et al. *Nat Genet* 2007;39(12):1477-82. PMID: 17982456. PMCID: PMC2652744.
  Wellcome Trust Case Control Consortium, et al. *Nat Genet* 2007;39(11):1329-37. PMID: 17952073. PMCID: PMC2680141.
  Nath SK, et al. *Nat Genet* 2008;40(2):152-4. PMID: 18204448.
  Hom G, et al. *N Engl J Med* 2008;358(9):900-9. PMID: 18204098.
  Liu Y, et al. *PLoS Genet* 2008;4(3):e1000041. PMID: 18369459. PMCID: PMC2274885.
  Nair RP, et al. *Nat Genet* 2009;41(2):199-204. PMID: 19169254. PMCID: PMC2745122.
  Barrett JC, et al. *Nat Genet* 2009;41:703-707. PMID: 19430480. PMCID: PMC2889014.
- For more information, see http://www.niams.nih.gov/News_and_Events/Press_Releases/2007/10_04.asp
- For more information, see http://grants.nih.gov/grants/guide/pa-files/PAR-09-135.html
- For more information, see http://www.niams.nih.gov/News_and_Events/Spotlight_on_Research/2008/Ankyl_Spond_gene.asp
- For more information, see http://grants.nih.gov/grants/guide/pa-files/PAR-08-123.html
- For more information, see http://www.nature.com/ng/journal/v41/n6/abs/ng.381.html
- This example also appears in Chapter 2: *Autoimmune Diseases* and Chapter 3: *Genomics*
- (E/I) (**NIAMS**, NCRR, NHGRI, NHLBI, NIAID, NICHD, NIDA, NIDCR, NIDDK)

The National Ophthalmic Disease Genotyping and Phenotyping Network (eyeGENE): Over the past 20 years, vision researchers have been remarkably successful in identifying the genetic basis of eye disease. More than 400 disease genes causing a wide range of eye diseases have been isolated, revealing unimagined complexity. Although some gene mutations lead to clearly defined clinical characteristics, or phenotypes, many other mutations are clinically indistinguishable from one another. Matching genetic testing with the disease phenotype will help to resolve this complexity and allow clinicians to diagnose specific diseases more accurately. However, commercial testing for rare eye diseases is limited. eyeGENE will expand the Nation's capacity to genotype patients with eye disease, thus improving patients' knowledge of their condition and the potential to personalize treatment eventually. From a research standpoint, patient DNA samples are invaluable for molecular studies, and patient registries are critical for patient recruitment for clinical trials. To address these needs, the NIH created eyeGENE, a partnership between government, health care providers, private industry, and scientists to broaden research resources and increase patient accessibility to diagnostic genetic testing. eyeGENE will provide researchers with patient genotype and phenotype data to elucidate ophthalmic disease genes and genetic modifiers, and enhance future enrollment of subjects in clinical trials.

- Brooks BP, et al. *Arch Ophthalmol* 2008;126(3)424-5. PMID: 18332328.
- For more information, see http://www.nei.nih.gov/resources/eyegene.asp
- (E/I) (**NEI**)

NINDS Human Genetics Repository: In 2002, NINDS established the Human Genetics Repository to collect, store, characterize, and distribute DNA samples and cell lines and standardized clinical data for the research community. By June 2009, the repository held material from 27,166 subjects, including those with cerebrovascular disease (8,625), epilepsy (1,356), Parkinson's disease (5,700), motor neuron diseases such as amyotrophic lateral sclerosis, also known as Lou Gehrig's disease, (2,631), and Tourette Syndrome (1,185), as well as control samples (6,162). The ethnically diverse collection represents populations from the United States and several other countries. Investigators have submitted or published more than 100 scientific articles based on data from this resource, and technological advances allowing whole genome screening for disease genes also have

enhanced its value.

- For more information, see http://ccr.coriell.org/Sections/Collections/NINDS/?SsId=10
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System*
- (E, I) (**NINDS**)

Dietary Supplement Ingredient Database (DSID): Working with the Nutrient Data Laboratory, Beltsville Human Nutrition Research Center, which is part of the USDA Agricultural Research Service, and other Federal agencies, NIH developed the DSID to estimate levels of ingredients in dietary supplement products. The main features of the database include data files, a research summary, and an adult multivitamin/minerals calculator. Since more than half of American adults report taking a dietary supplement, the estimates in the DISD will improve assessment of total nutrient intake from foods and supplements.

- For more information, see http://dietarysupplementdatabase.usda.nih.gov
- (E) (**ODP/ODS**)

Dietary Supplement Labels Database (DSLD): NIH is developing a comprehensive information resource on dietary supplements labels. The current database includes information from the labels of approximately 4,000 dietary supplement products in the marketplace, including vitamins, minerals, herbs or other botanicals, amino acids, and other specialty supplements. Ingredients of dietary supplements in this database are linked to other NIH databases such as MedlinePlus® to allow users to investigate the dietary ingredients and view biomedical literature pertaining to them. NIH is piloting the development of a full-scale application that includes label information on virtually all dietary supplements sold in the United States. The future DSLD will provide comprehensive label information in a format that is user-friendly for both consumers and researchers. The information included in the database will be determined by Federal and stakeholder user groups.

- For more information, see http://dietarysupplements.nlm.nih.gov
- (E) (**NLM**, ODP/ODS)

Disease Registries and Surveillance Systems: Tracking and Monitoring Disease

Seeking Solutions for People with Sjogren's Syndrome: Sjogren's syndrome is one of the most prevalent autoimmune disorders, affecting as many as 4 million people in the United States. Nine out of 10 patients affected are female. It is an autoimmune disease that progressively destroys salivary and lachrymal glands. The most common symptoms include dry eyes, dry mouth, fatigue, and musculoskeletal pain. A significant roadblock for moving discoveries ahead in the field of Sjogren's syndrome is the lack of data and biospecimens available for research. Recognizing the problem, NIH spearheaded an effort to establish Sjogren's patient registries at two extramural institutions as well as through its own intramural program. These groups are working together to generate and share with the general research community the genome-wide genotyping data and clinical information from the cohorts enrolled through these efforts. This resource should jumpstart efforts to understand genetic contributions to Sjogren's syndrome and the etiologic overlap with related autoimmune conditions such as lupus and rheumatoid arthritis. In addition to participating in the patient registry and genotyping efforts described above, the Sjogren's Syndrome Clinic, located in the NIH CC, collects systematic clinical and laboratory data on the Sjogren's syndrome (and salivary dysfunction) population. Gene

therapy and bioengineering hold promise for the repair or even replacement of salivary glands ravaged by Sjogren's syndrome. More than 300 patient visits occur annually, and the clinic is expanding its patient recruitment to accelerate the conduct of clinical trials that might shed light on this disorder.

- Korman BD, et al. *Genes Immun* 2008;9(3):267-70. PMID: 18273036.
  Roescher N, et al. *Oral Dis* 2009;15(8):519-26. PMID: 19519622. PMCID: PMC2762015.
  Nikolov NP, Illei GG. *Curr Opin Rheumatol* 2009;21(5):465-70. PMID: 19568172. PMCID: PMC2766246.
- For more information, see http://www.sjogrens.org/
- This example also appears in Chapter 2: *Autoimmune Diseases* and Chapter 3: *Genomics*
- (E/I) (**NIDCR**, CC, ORWH)

International Epidemiologic Databases to Evaluate AIDS (IeDEA): The goal of the IeDEA program is to conduct analyses based on comparable data from multiple regions and studies. This initiative has established international regional centers for the collection and harmonization of data and has created an international research consortium to address unique and evolving research questions in HIV/AIDS currently unanswerable by single cohorts. High-quality data are being collected by researchers throughout the world. This initiative provides a means to establish and implement methodology to pool the collected data effectively—thus providing a cost-effective means of generating large data sets to address the high-priority research questions. Combination of data collected under various protocols frequently is very difficult and not as efficient as the collection of predetermined and standardized data elements. By developing a proactive mechanism for the collection of key variables, this initiative will enhance the quality cost effectiveness and speed of HIV/AIDS research. Participating regions include Canada and the United States, the Caribbean and Central and South America, Asia and Australia (excluding China), West Africa, Central Africa, East Africa, and Southern Africa.

- For more information, see http://www3.niaid.nih.gov/about/organization/daids/daidsepi.htm
- This example also appears in Chapter 2: *Infectious Diseases and Biodefense*
- (E) (**NIAID**, NCI, NICHD)

Environmental Polymorphisms Registry: NIH, in collaboration with the University of North Carolina's General Clinical Research Center, has launched a large volunteer DNA banking project named the Environmental Polymorphisms Registry (EPR). The goal of the EPR is to collect DNA samples from 20,000 individuals in the greater Research Triangle Park region of North Carolina through local health care systems, study drives, health fairs, and other means. This area has a diverse population varying in age, ethnicity, economic and educational backgrounds, and health status. The EPR offers a valuable resource for human genomic studies, especially when compared to anonymous DNA registries. It was designed for scientists to screen for functionally significant alleles and to identify subpopulations of individuals with shared genotypes, and then correlate their genotypes with their phenotypes in a process known as "recruit-by-genotype." The value of the EPR lies in the ability to identify and then re-contact subjects with potentially significant polymorphisms for further study. A unique feature of the EPR is that two distinct populations are solicited, an apparently healthy population recruited from the general population as well as a clinic population recruited from various clinics and hospitals in the area. Individuals in the clinic population have a wide array of medical conditions, and their inclusion in the EPR increases the likelihood of identifying subjects with both the genotypes and phenotypes of interest. These aspects of the EPR give scientists more flexibility in designing follow-up studies while reducing the ascertainment bias that can occur in genetic epidemiology studies when subjects are recruited based on phenotype.

- This example also appears in Chapter 3: *Epidemiological and Longitudinal Studies*
- (E/I) (**NIEHS**)

Surveillance, Epidemiology, and End Results (SEER): The SEER program provides essential data that support cancer research across NIH and collaborating agencies and organizations in the United States and around the world. SEER covers approximately 26 percent of the U.S. population, with information in its database on more than 5.7 million cancer cases. SEER registries routinely collect data on patient demographics, primary tumor site, morphology, extent of disease at diagnosis, and first course of treatment. All patients are followed annually for vital status and compilation of survival data. The SEER Program is the only comprehensive source of population-based data in the United States that includes stage of cancer at the time of diagnosis and survival rates by stage. It is the only population-based source of long-term incidence and survival data, having a 35-year history in most of its registries. SEER provides source data for the American Cancer Society Facts & Figures and the Annual Report to the Nation on the Status of Cancer. SEER is one of the most fundamental contributors to the cancer research infrastructure, adding more than 380,000 cases each year. The program sets national benchmarks for incidence and survival rates and is the primary source of reports on cancer death rates. The size of the database allows for analysis of rare cancers and cancer heterogeneity at both the tumor and patient level. The SEER database also includes prevalence information on the 11.4 million cancer survivors in the United States, allowing analysis by age and cancer site as well as time elapsed since diagnosis. There are more than 2,000 agreements executed annually for the public-use data and more than 3 million hits per month on the SEER Internet homepage.

- For more information, see http://seer.cancer.gov
- This example also appears in Chapter 2: *Cancer*
- (E) (**NCI**)

Cancer Control P.L.A.N.E.T: The Cancer Control P.L.A.N.E.T. (Plan, Link, Act, Network with Evidence-based Tools) Web portal was launched collaboratively in 2003 by NIH, Agency for Healthcare Research and Quality, American Cancer Society, Centers for Disease Control and Prevention, Commission on Cancer, and Substance Abuse and Mental Health Services Administration. The portal now has been expanded, in collaboration with the Surveillance Action Group of the Canadian Partnership Against Cancer, to include Cancer Control P.L.A.N.E.T. Canada. The Canadian site follows the same design as the U.S. site, while engaging Canadian cancer control practitioners and researchers in usability testing to ensure that the Canadian site meets their needs. Both the Canadian and U.S. sites provide a single point of access to high-quality tools and resources from multiple national organizations that can be used to design, implement, and evaluate evidence-based cancer control plans and programs. They guide local programs to resources that help them determine cancer risk and cancer burden in their geographic areas. They also help identify potential partners and provide online resources for interpreting research findings and recommendations and accessing products and guidelines for planning and evaluation.

- For more information, see http://cancercontrolplanet.cancer.gov
- This example also appears in Chapter 2: *Cancer*
- (E) (**NCI**)

A Look at Drug Abuse Trends: Local to International: Two major systems of data collection are helping to identify substance abuse trends locally, nationally, and internationally: Monitoring the

Future Survey (MTF) and the Community Epidemiology Work Group (CEWG). Both help to surface emerging drug abuse trends among adolescents and other populations, and guide responsive national and global prevention efforts. The MTF project, begun in 1975, has many purposes, the primary one being to track trends in substance use, attitudes, and beliefs among adolescents and young adults. The survey findings also have been used by the President's Office of National Drug Control Policy to monitor progress toward national health goals. The MTF project includes both cross-sectional and longitudinal formats—the former given annually to 8th, 10th, and 12th graders to see how answers change over time, and the latter given every 2 years (until age 30), then every 5 years to follow up on a randomly selected sample from each senior class. CEWG, established in 1976, provides both national and international information about drug abuse trends through a network of researchers from different geographic areas. Regular meetings feature presentations on selected topics, as well as those offering international perspectives on drug abuse patterns and trends. CEWG findings reported in 2008 and 2009 show decreases in methamphetamine indicators (e.g., treatment admissions), suggesting that the problems that had escalated in the first half of the decade may have stabilized or declined. Development of a Latin American Epidemiology Network is underway. NIH also has provided technical consultation for the planning and establishment of an Asian multicity epidemiological network on drug abuse.

- For more information, see http://www.monitoringthefuture.org/
- For more information, see http://www.drugabuse.gov/about/organization/CEWG/CEWGHome.html
- This example also appears in Chapter 2: *Minority Health and Health Disparities* and Chapter 3: *Epidemiological and Longitudinal Studies*
- (E) (**NIDA**)

Alcohol Policy Information System: Public policies that affect alcohol consumption and related behaviors can influence a range of health and social outcomes. The NIH has developed the Alcohol Policy Information System (APIS) to provide authoritative and detailed information on alcohol-related public policies in the United States at both the State and Federal levels. Intended primarily as a tool for researchers, the APIS website (http://alcoholpolicy.niaaa.nih.gov), posted in June 2003, features compilations and analyses of alcohol-related statutes and regulations. APIS is designed to simplify the process of ascertaining the state of the law for studies on the effects and effectiveness of alcohol-related policies. APIS currently provides information on 30 specific policy topics, including summary descriptions, maps, detailed comparison tables, and the specific dates on which provisions became or ceased to be effective. For most policy topics, APIS coverage begins as early as January 1, 1998, and extends through September 18, 2008. NIH issued program announcements for alcohol policy research that use APIS in 2007.

- Fell J, et al. *Alcohol Clin Exp Res* 2009;33(7):1208-1219.PMID: 19389192.
  Wagenaar AC, et al. *Addiction* 2009;104(2):179-90. PMID: 19149811.
- For more information, see http://www.alcoholpolicy.niaaa.nih.gov
- (E) (**NIAAA**)

Standardized Vocabularies, Data Protocols, and Tools

Health IT Standards and Electronic Health Records: NIH researchers are engaged in developing Next Generation electronic health records (EHRs) with advanced decision-support capabilities to facilitate patient-centered care, clinical research, and public health. As the central

coordinating body for clinical terminology standards within HHS, NIH works closely with the Office of the National Coordinator for Health Information Technology (ONC) to support nationwide implementation of an interoperable health information technology infrastructure. NIH develops or licenses key clinical terminologies that are designated as standards for U.S. health information exchange. The Unified Medical Language System Metathesaurus, with more than 8.1 million concept names from more than 125 vocabularies, is a distribution mechanism for standard code sets and vocabularies used in health data systems. NIH also produces RxNorm, a standard clinical drug vocabulary; supports the LOINC nomenclature for laboratory tests and patient observations; and collaborates with the International Health Terminology Standards Development Organisation to promote international adoption of the SNOMED CT clinical terminology. In FY 2009, NIH released the first version of the CORE Problem List Subset of SNOMED CT, designed to facilitate coding of problem list data in EHRs by mapping frequently used terms from seven large-scale health care institutions to corresponding SNOMED CT concepts. The Newborn Screening Codes and Terminology Guide, a Web portal to support more effective use of newborn screening laboratory test information, was created in FY 2009 in collaboration with ONC, the Health Resources and Services Administration, and newborn screening organizations.

- For more information, see http://www.nlm.nih.gov/research/umls
- This example also appears in Chapter 3: *Technology Development*
- (I) (**NLM**)

Health Information Technology: Health information technology research that enables the integration of clinical data and medical image diagnostic and treatment data with the patient's medical history in a comprehensive electronic medical record will improve clinical decision-making. The ability to connect and exchange diagnostic information and medical images between health care providers, clinics, and hospitals will help provide the timely information that is needed for effective health care and will help reduce unnecessary, excessive, and duplicative procedures. A patient-centered approach to comprehensive electronic health records will allow patients access to their health information. This will enable patients to play an active role in their own wellness by enabling them to ask knowledgeable questions about treatment options. Additionally, patients also are empowered to provide this information to any and all health care providers as needed, independent of their location or where the medical data was created or stored. NIH supports research in new methods and technologies to address issues such as: interoperability of data systems, compatibility of computer software across medical institutions, security of data during transmission, HIPAA compliance, availability of affordable data systems for patient care providers, and integration of medical decision-support information in medical data systems.

- This example also appears in Chapter 3: *Technology Development*
- (E) (**NIBIB**, NLM)

Discovery Initiative for Entrez Databases: The Discovery Initiative aims to maximize the utility of NIH biomedical data resources by better exploiting their inter-linkages. For example, a PubChem record on a chemical structure might link to records for similar proteins, related protein structures, and relevant journal articles. Such linkages provide users with tremendous opportunities for exploration and scientific discovery, but are underutilized currently. The Discovery Initiative aims to improve the retrieval and presentation of results so that users are drawn more readily to related data that could lead to serendipitous discoveries. Improvements have been made in the search interface through the use of "sensors" that can detect certain categories of search terms automatically, such as genes or drugs, and then direct the user's attention to resources that may augment the original search. Through these linkages, users are better able to traverse the 40-plus databases in the Entrez network, ranging from

topics such as human genetic disorders and genome projects to cancer chromosomes and protein structure.

- National Center for Biotechnology Information. Featured Resource: Improvements to NCBI Services Promote Discovery. *NCBI News* 2009;February:1.
- (I) (**NLM**)

Collective Intelligence for Knowledge Discovery: NIH has started a new NIH initiative in collective intelligence. The goal is to create deep repositories of knowledge backed by controlled vocabularies or ontologies, and to create or enhance semantically interoperable applications capable of discovering knowledge hidden within these repositories. Current applications such as the Human Salivary Proteome Annotation System, the Common Assay Reporting System, and the caBIG Protocol Lifecycle Tracking Tool are among the initial steps of a knowledge infrastructure. These applications harvest the collective knowledge of targeted scientific communities to store protocols, data, and results. Other tools developed for this initiative (e.g., the context-sensitive text mining system for identification of high-risk, high-reward research) use statistical natural language processing to discover new knowledge, such as, whether in peer review, an application for funding was considered high-risk and high-reward. Additional pilot studies are evaluating computational linguistics and knowledge management tools for biomedical and clinical informatics, portfolio analysis, systems biology, proteomics, genomics, and knowledge representation paradigms. The collective-intelligence initiative will lead to a knowledge infrastructure that can shift the paradigms of data re-use and knowledge discovery dramatically.

- This example also appears in Chapter 3: *Molecular Biology and Basic Research*
- (I) (**CIT**, CC, NCI, NHGRI, NIDCR, NIMH, OD)

NIH Federated Identity Service: NIH Federated Identity Service enables people from institutions external to NIH to collaborate by allowing them to use their user name or password from their home organization to access authorized NIH systems. Federated Identity maintains user privacy by keeping the users' credentialing process within their home organization while enabling more seamless collaborations and transactions between federated organizations that can trust each other's identity authentication. Federated Identity Service currently federates with more than 22 institutions, including its sister operational division, the FDA, and universities such as Johns Hopkins, Duke, and Ohio State. NIH made its Clinical & Translational Science Awards (CTSA) Wiki one of the first NIH systems to be federated with nongovernment institutions. Federated Identity facilitates access to the wiki—an online, authorized access, collaborative work environment for members of the CTSA Consortium, which currently supports 1,200 members at 38 universities. The CTSA program reports that 15 of their awarded institutes actively federate with NIH. Other services accepting external credentials through Federated Identity include the website NCRR Annual Progress Report Scientific Information System, FDA ITAS Time and Attendance, NLM NCBI SharePoint for Genome Research, and the Salivary Proteome Wiki. In addition to accepting external accounts, NIH users may use their username and password to access such diverse services as GovTrip, the CTSA Indiana University Wiki, and the Genome Browser at University of California Santa Cruz.

- (I) (**CIT**)

Large-Scale Informatics Infrastructure

The Cancer Biomedical Informatics Grid® (caBIG®): The caBIG® initiative connects researchers and institutions to enable collaborative research and personalized, evidence-based care. More than 1,500 individuals representing more than 450 government, academic, advocacy, and commercial organizations have collaborated to develop a standards-based grid infrastructure (caGrid) and a diverse collection of interoperable software tools, enabling basic and clinical researchers to speed the translation of information from bench to bedside. Forty-nine of the 65 NCI-designated Cancer Centers and 8 of 10 organizations of the NCI Community Cancer Centers Program are actively deploying caBIG® tools and infrastructure in support of their research efforts. Additionally, caBIG® technology is adapted to power noncancer research initiatives such as the CardioVascular Research Grid. Ongoing collaborations with research and bioinformation organizations in the United Kingdom, China, and India are driving international adoption of caBIG® resources. The caBIG® infrastructure also supports a new health care ecosystem, BIG Health™, in collaboration with various stakeholders in biomedicine (e.g., government, academia, industry, nonprofits, and consumers) in a novel organizational framework to demonstrate the feasibility and benefits of personalized medicine. BIG Health™ will provide the foundation for a new approach in which clinical care, clinical research, and scientific discovery are linked.

- For more information, see http://cabig.cancer.gov
- For more information, see http://bighealthconsortium.org/
- This example also appears in Chapter 2: *Cancer* and Chapter 3: *Technology Development*
- (E/I) (**NCI**)

Biomedical Informatics Research Network (BIRN): Modern biomedical research generates vast amounts of diverse and complex data. Increasingly, these data are acquired in digital form, allowing sophisticated and powerful computational and informatics tools to help scientists organize, store, query, mine, analyze, view, and, in general, make better use and sense of their data. Moreover, the digital form of these data and tools makes it possible for them to be shared easily and widely across the research community at large. NIH has supported development of the BIRN infrastructure to share data and tools by federating new software tools or using the infrastructure to federate significant datasets. BIRN fosters large-scale collaborations by using the capabilities of the emerging national cyberinfrastructure. In FY 2009, the BIRN Coordinating Center transitioned to a new home at the University of Southern California. The new BIRN Coordinating Center uses grid computing technology to create a virtual organization for basic and clinical science investigators across the network. In addition, a new BIRN Community Service (U24) grant was awarded to help expand the BIRN user community to researchers and clinicians beyond the neuroscience and imaging fields.

- For more information, see http://www.ncrr.nih.gov/birn
- For more information, see http://www.nbirn.net
- This example also appears in Chapter 3: *Technology Development*
- (E) (**NCRR**)

A Clearinghouse for Neuroimaging Informatics Tools and Resources: Many neuroimaging tools and databases are underutilized because they cannot be found easily, are not user-friendly, or are not easily adoptable or adaptable. In an effort to promote the enhancement, adoption, distribution, and evolution of neuroimaging informatics tools and resources, the NIH Blueprint for Neuroscience Research has launched the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC). Examples of included tools are: image segmentation, image registration, image processing pipelines, statistical analysis packages, spatial alignment and normalization algorithms, and data format

translators. Resources include: well-characterized test datasets, data formats, and ontologies. Since the first release in October 2007, the clearinghouse website, or NITRC, has become host to 180 tools and resources, with a community of 13,602 unique visitors who downloaded NITRC tools and resources, and 7,000 unique visitors per month, more than 954 of which are registered users (11 percent non-English speaking). The hits to the site have reached 15,635,019/month. Since its inception, more than 50,000 software files have been downloaded. More than 53 percent of the tools on NITRC had not been shared online previously but now are available to the community. In 2009, the NITRC project won the first place of Excellence.gov awards, the largest Federal government award program to recognize the very best in government IT programs, among 61 competitors. Through the initiative, nearly 40 awards have been made to neuroimaging tools and resource developers to enhance the accessibility, interoperability, and adoptability of their existing tools and resources.

- Ardekani BA, Bachman AH. *Neuroimage* 2009;46(3):677-82. PMID: 19264138. PMCID: PMC2674131.
- For more information, see http://www.nitrc.org/
- For more information, see http://neuroscienceblueprint.nih.gov/
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System* and Chapter 3: *Technology Development*
- (E) (**NIH Blueprint**, NCCAM, NCRR, NEI, NIA, NIAAA, NIBIB, NICHD, NIDA, NIDCD, NIDCR, NIEHS, NIGMS, NIMH, NINDS, NINR, OBSSR)

National Centers for Biomedical Computing: There are seven NIH Roadmap National Centers for Biomedical Computing (NCBC). Funded as cooperative agreements, these centers collectively cover broad areas of neuroinformatics, functional genomics, image post processing, multiscale modeling, cellular pathways, semantic data integration and ontologies, information networks, cellular networks and pathways, clinical informatics, disease-gene-environment analysis, and clinical decisions support.

- For more information, see http://ncbcs.org/
- This example also appears in Chapter 3: *Molecular Biology and Basic Research* and Chapter 3: *Technology Development*
- (E) (**NIGMS**, Common Fund - all ICs participate)

Disaster Information Services: A Disaster Information Management Research Center was established in FY 2008 with the aim to facilitate access to disaster information, promote more effective use of libraries and disaster information specialists in disaster management efforts, and ensure uninterrupted access to critical health information resources when disasters occur. A disaster information website provides access to a broad range of emergency preparedness and response information. The Center also collaborates with the Navy National Medical Center, Suburban Hospital, Johns Hopkins Medicine, and NIH CC in the Bethesda Hospital Emergency Preparedness Partnership to provide backup communication systems and develop tools for patient tracking, information sharing and access, and responder training and to serve as a model for hospitals across the Nation. NIH also develops advanced information services and tools to assist emergency responders when disaster strikes. WISER (Wireless Information System for Emergency Responders) was developed for use during hazardous materials incidents and is available on the Internet or for downloading onto PDAs and PCs. Usage continues to grow, with more than 47,000 downloads onto PDAs in FY 2008. Radiation Event Medical Management (REMM) is a downloadable toolkit for use by health care providers during a mass casualty radiation event, with a version for mobile platforms released in FY 2008. Developed in collaboration with the HHS Office of Public Health Preparedness, REMM includes procedures for

diagnosis and management of radiation contamination and exposure, guidance for use of radiation medical countermeasures, among other features to facilitate medical responses to radiation emergencies.

- For more information, see http://disasterinfo.nlm.nih.gov
- For more information, see http://wiser.nlm.nih.gov
- For more information, see http://remm.nlm.gov
- This example also appears in Chapter 3: *Health Communication and Information Campaigns and Clearinghouses*
- (I) (**NLM**)

Health Care Delivery Consortia to Facilitate Discovery and Improve Quality of Cancer Care: The purpose of the Cancer Research Network (CRN) is to enhance research on cancer epidemiology, prevention, early detection, and control in the context of health care delivery systems. CRN combines established research groups affiliated with 14 health care delivery organizations that provide comprehensive care to a racially and ethnically diverse population of nearly 11 million individuals. CRN has developed strong research capabilities in several areas: developing and applying innovative methods to collect and interpret data from both conventional and electronic medical records systems; assembling large samples of patients with documentation of patient characteristics and longitudinal data on receipt of health services and clinical and quality-of-life outcomes; collecting and integrating complex data from patients, providers, and organizations to examine issues in health care delivery from multiple perspectives; quantifying the effect of key factors in the delivery process that may determine quality and outcomes of care; and conducting studies on behavioral and systems-based interventions to improve the delivery of care in community-based health care delivery systems. The Breast Cancer Surveillance Consortium (BCSC) is a research resource for studies designed to assess the delivery and quality of breast cancer screening and related patient outcomes in the United States. The BCSC is a collaborative network of seven mammography registries with linkages to tumor and/or pathology registries. The Consortium's database contains information on 7,521,000 mammographic examinations, 2,017,869 women, and 86,700 cancer cases.

- For more information, see http://crn.cancer.gov
- For more information, see http://breastscreening.cancer.gov/
- This example also appears in Chapter 2: *Cancer*, Chapter 3: *Epidemiological and Longitudinal Studies* and Chapter 3: *Clinical and Translational Research*
- (I) (**NCI**)

CISNET—A Resource for Comparative Effectiveness Research: The Cancer Intervention and Surveillance Modeling Network (CISNET) represents a quantum leap forward in the practice of modeling to inform clinical and policy decisions. While contemporary science has enabled the collection and analysis of health-related data from numerous sectors, enormous challenges remain to integrate various sources of information into optimal decision-making tools to inform public policy. Collaborative work on key questions promotes efficient collecting and sharing of the most important data and critical evaluation of the strengths and weaknesses of each resource. Providing results from a range of models, rather than a single estimate from one model, brings credibility to the process and reassures policymakers that the results are reproducible. CISNET is a consortium of NIH-sponsored investigators who use modeling to improve understanding of the impact of cancer control interventions (e.g., prevention, screening, and treatment) on incidence and mortality trends. The consortium's work informs clinical practice and recommended guidelines by synthesizing existing information to model gaps in available knowledge. CISNET provides a suite of models that are poised to determine the most efficient

and cost-effective strategies for implementing technologies in the population. Four groups of grantees focus on breast, prostate, colorectal, and lung cancers using statistical simulation and other modeling approaches. Their models incorporate data from randomized controlled trials, meta-analyses, observational studies, epidemiological studies, national surveys, and studies of practice patterns to evaluate the past and potential future impact of these interventions.

- For more information, see http://cisnet.cancer.gov/
- This example also appears in Chapter 2: *Cancer* and Chapter 3: *Molecular Biology and Basic Research*
- (E/I) (**NCI**)

NIH Biowulf Cluster Enables Large-Scale Biomedical Research: The Biowulf cluster provides NIH researchers with a world-class supercomputer that enables the conduct of large-scale biomedical computational projects, allowing scientific research that otherwise would not be possible. Biowulf comprises more than 6,000 interconnected processors operating cooperatively to solve such diverse problems as: identifying genotype patterns of variation across worldwide human populations; validating algorithms used in computer-aided detection of colon polyps ("Virtual Colonoscopy"); computing the molecular structures of viruses such as HIV using 3D electron microscopy; facilitating whole-genome assembly and genome-wide association studies resulting from next-generation DNA-sequencers; and, as part of the NIH Roadmap Initiative for Molecular Libraries, generating conformation ensembles for 25 million chemical structures. In 2008-2009, more than 105 scientific papers published by NIH intramural scientists cited the use of Biowulf as a computational resource.

- This example also appears in Chapter 3: *Technology Development*
- (I) (**CIT**)

Biomedical Informatics Research and Training

Informatics Research Training Programs: Exploiting the potential of information technology to augment health care, biomedical research, and education requires investigators who understand biomedicine as well as knowledge representation and decision support. NLM is the principal source of extramural funding for research training in the fields of biomedical informatics, supporting approximately 270 trainees at 18 institutional training programs throughout the country. NLM also provides intramural informatics research training opportunities for another 70 students, postdoctorates, and visiting scientists, as well as training and career development fellowships for health science librarians on the NIH campus and at academic health sciences centers across the country. Collectively, NLM's research training programs encompass health care informatics, bioinformatics, clinical research translational informatics, and public health informatics. Recent highlights and developments in informatics training include:

- A congressional supplemental appropriation for FY 2008 allowed NIH to add 26 NLM training slots.
- A Diversity Short-Term Trainee Program was implemented to improve the diversity of informatics trainees, with funding for 18 trainees at 7 training programs.
- Funds from the American Reinvestment and Recovery Act were committed to support an additional 56 2-year slots at 10 of its informatics training programs.
- A new Clinical Informatics Postdoctoral Fellowship was established to attract young physicians to NIH to pursue research in informatics.

- For more information, see http://www.nlm.nih.gov/training.html
- This example also appears in Chapter 3: *Research Training and Career Development*
- (E/I) (**NLM**) (ARRA)